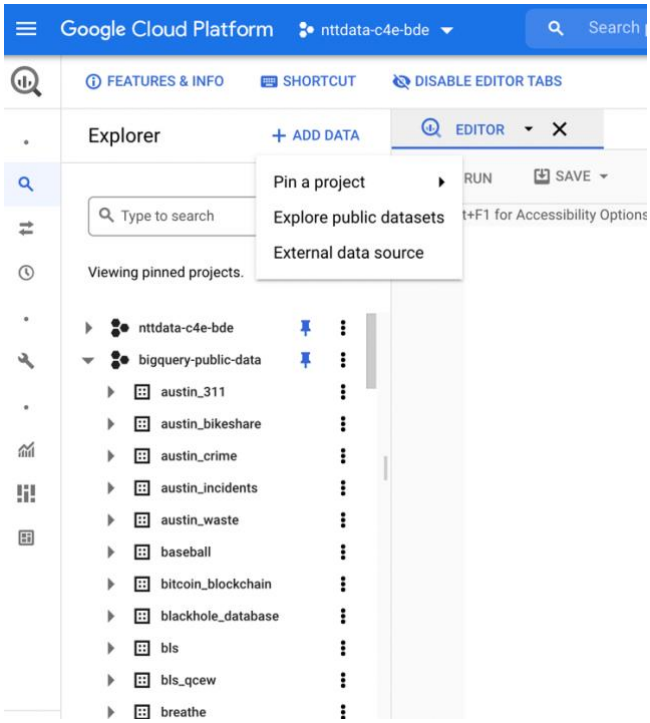
Lab n. 2: Use of BigQuery

This Lab will focus on using BigQuery and the **public datasets** available on GCP.
To show the public datasets you can click on ADD DATA →Explore public datasets, under BigQuery panel.



In this lab, for each exercise, you will execute queries, then you will save these in a public repo on GitHub.

The queries will be saved in .sql files which must have the following format:
name_surname_LAB2_Exercise_<x>.sql
Example:
Will_Smith_LAB2_Exercise_1.sql

Pay attention: before running each query, always check how much data you move: each of you will have “only” 100GB a day available.

Enjoy!

Exercise n1.
Given the shared file top_4000_movies_data.csv

1. Create a BigQuery table “Movie”
2. Find the top 10 highest budget films, year by year, from 2016 to 2020.

Output example:

Row	year	Movie_Title	Production_Budget	rank
1	2020	Tenet	205000000	1
2	2020	Onward	200000000	2
3	2020	Wonder Woman 1984	200000000	2
4	2020	Mulan	200000000	2
5	2020	Dolittle	175000000	5
6	2020	The Call of the Wild	125000000	6
7	2020	Artemis Fowl	100000000	7
8	2020	Bad Boys For Life	90000000	8
9	2020	Sonic The Hedgehog	90000000	8
10	2020	Birds of Prey (And the Fantabulous Emancipation...	82000000	10
11	2019	Avengers: Endgame	400000000	1
12	2019	Star Wars: The Rise of Skywalker	275000000	2
13	2019	The Lion King	260000000	3
14	2019	Toy Story 4	200000000	4
15	2019	Fast & Furious Presents: Hobbs & Shaw	200000000	4
16	2019	Dark Phoenix	200000000	4

Exercise n. 2
Show a flat result of the pages visited on 1st August 2017

Public Dataset: bigquery-public-data.google_analytics_sample
Table involved: ga_sessions_20170801

Output example:

Row	visitId	visitStartTime	pageTitle	pagePath
1	1501583974	1501583974	Electronics | Google Merchandise Store	/google+redesign/electronics
2	1501616585	1501616585	Men's Outerwear | Apparel | Google Merchandise Store	/google+redesign/apparel/mens/mens+outerwear
3	1501583344	1501583344	Android | Shop by Brand | Google Merchandise Store	/google+redesign/shop+by+brand/android
4	1501573386	1501573386	Power & Chargers | Electronics | Google Merchandise Store	/google+redesign/electronics/power
5	1501651467	1501651467	Men's Apparel | Google Merchandise Store	/google+redesign/apparel/mens

Exercise n. 3
Find how many times a page it was visited in July 2017. Group by date the result

Public Dataset: bigquery-public-data.google_analytics_sample
Table involved: ga_sessions_201707* (wildcard)

Output example:

Row	date	pagePath	counter
1	2017-07-01	/home	1005
2	2017-07-01	/google+redesign/shop+by+brand/youtube	764
3	2017-07-01	/google+redesign/shop+by+brand/waze+baby+on+board+window+decal.axd	298

Exercise n. 4

1. Find the first 3 most used pairs of OS and Browser from mobile devices for each available country on 1st August 2017
2. Produce a new table with the query result

Public Dataset: bigquery-public-data.google_analytics_sample
Table involved: ga_sessions_20170801
Tip: ARRAY_AGG functions can be useful
(see: https://cloud.google.com/bigquery/docs/reference/standard-sql/aggregate_functions)

Output example:

Row	country	country_rank.operatingSystem	country_rank.browser	country_rank.rank
1	Algeria	Android	Chrome	1
2	Argentina	Android	Chrome	1
3	Australia	iOS	Chrome	1
		Android	Opera	2
		Android	Chrome	3
4	Australia	iOS	Chrome	1
		Android	Opera	2
		Android	Chrome	3
5	Australia	iOS	Chrome	1
		Android	Opera	2
		Android	Chrome	3
6	Austria	Windows	Internet Explorer	1
		Android	Chrome	2
7	Austria	Windows	Internet Explorer	1
		Android	Chrome	2
8	Bahrain	iOS	Safari	1
9	Bangladesh	Android	Nokia Browser	1

Exercise n. 5

Find the top 10 users (the id) who answered the most questions.

Public Dataset: bigquery-public-data.stackoverflow
Table involved: posts_answers, users

Output Example:

Row	id_user	count
1	13249	5892
2	17034	4949
3	22656	4920
4	29407	4598
5	157882	4385
6	20862	4248
7	187606	4000
8	34397	3963
9	70604	3700
10	61974	3334

Exercise n. 6
Find the top 10 StackOverflow users (the id) by accepted responses on 2010 posts

Public Dataset: bigquery-public-data.stackoverflow
Table involved: stackoverflow_posts, posts_answers, users

Output Example:

Row	id_user	count
1	13249	3865
2	22656	2923
3	17034	2745
4	157882	2438
5	29407	2272
6	70604	2004
7	34397	1852
8	187606	1844
9	61974	1713
10	115145	1561

Exercise n. 7
Find the most popular programming language

Public Dataset: bigquery-public-data.github_repos
Table involved: languages

Output Example:

Row	language_name	count
1	JavaScript	1100360

Exercise n. 8
Find the top 10 committers in 2016 on Github repositories that uses the Java language

Public Dataset: bigquery-public-data.github_repos
Tables involved: languages, sample_commits

Output Example:

Row	name	count
1	TensorFlower Gardener	2449
2	Benjamin Pasero	1127
3	Vijay Vasudevan	950
4	Joao Moreno	755
5	Alex Dima	711
6	isidor	697
7	Johannes Rieken	637
8	Martin Aeschlimann	430
9	Martin Wicke	286
10	Daniel Imms	275

Exercise n. 9
Find for each commit on Github on a .c file of the Linux kernel, the previous and the next commit.

Public Dataset: bigquery-public-data.github_repos
Table involved: sample_commits
Tip: Lag and Lead functions can be useful
(see: https://cloud.google.com/bigquery/docs/reference/standard-sql/navigation_functions)

Output Example:

Row	repo_name	file	date	previous_commit	commit	next_commit
1	torvalds/linux	kernel/acct.c	2005-04-16 22:20:36 UTC	null	1da177e4c3f41524e886b7f1b8a0c1fc7321cac2	6c9c0b52b8c6b68b05bb06efd7079a8fc5e9ba60
2	torvalds/linux	kernel/acct.c	2005-09-07 23:57:31 UTC	1da177e4c3f41524e886b7f1b8a0c1fc7321cac2	6c9c0b52b8c6b68b05bb06efd7079a8fc5e9ba60	64e47488c913ac704d465a6af86a26786d1412a5
3	torvalds/linux	kernel/acct.c	2005-09-08 05:45:47 UTC	6c9c0b52b8c6b68b05bb06efd7079a8fc5e9ba60	64e47488c913ac704d465a6af86a26786d1412a5	5a2cec83a9bb1b4295aa8ab728fcb8ca1811a33c
4	torvalds/linux	kernel/acct.c	2005-09-08 09:37:58 UTC	64e47488c913ac704d465a6af86a26786d1412a5	5a2cec83a9bb1b4295aa8ab728fcb8ca1811a33c	c324b44c34050cf2a9b58830e11c974806bd85d8
5	torvalds/linux	kernel/acct.c	2005-09-08 09:39:55 UTC	5a2cec83a9bb1b4295aa8ab728fcb8ca1811a33c	c324b44c34050cf2a9b58830e11c974806bd85d8	142e27fc8a3619471669d6241784eec9167c47d1
6	torvalds/linux	kernel/acct.c	2005-09-08 09:41:28 UTC	c324b44c34050cf2a9b58830e11c974806bd85d8	142e27fc8a3619471669d6241784eec9167c47d1	1d6ae775d7a948c9575658eb41184fd2e506c0df
7	torvalds/linux	kernel/acct.c	2005-09-08 09:43:49 UTC	142e27fc8a3619471669d6241784eec9167c47d1	1d6ae775d7a948c9575658eb41184fd2e506c0df	344a076110f4ecb16ea6d286b63be696604982ed
8	torvalds/linux	kernel/acct.c	2005-09-08 21:27:13 UTC	1d6ae775d7a948c9575658eb41184fd2e506c0df	344a076110f4ecb16ea6d286b63be696604982ed	d344c5e0856ad03278d8700b503762dbc8b86e12
9	torvalds/linux	kernel/acct.c	2005-09-10 01:14:47 UTC	344a076110f4ecb16ea6d286b63be696604982ed	d344c5e0856ad03278d8700b503762dbc8b86e12	417ef531415c070926b071b75fd1c1ac4b6e2f7e
10	torvalds/linux	kernel/acct.c	2005-09-10 17:06:26 UTC	d344c5e0856ad03278d8700b503762dbc8b86e12	417ef531415c070926b071b75fd1c1ac4b6e2f7e	ad2c10f8f00d3fe2e37dd8a107e7cf4ac0459489
11	torvalds/linux	kernel/acct.c	2005-09-12 19:10:59 UTC	417ef531415c070926b071b75fd1c1ac4b6e2f7e	ad2c10f8f00d3fe2e37dd8a107e7cf4ac0459489	d58dde0f552a5c5c4485b962d8b6e9dd54fefb30
12	torvalds/linux	kernel/acct.c	2005-09-12 19:45:04 UTC	ad2c10f8f00d3fe2e37dd8a107e7cf4ac0459489	d58dde0f552a5c5c4485b962d8b6e9dd54fefb30	d7f6884ae0ae6e406ec3500fcde16e8f51642460
13	torvalds/linux	kernel/acct.c	2005-09-14 12:01:25 UTC	d58dde0f552a5c5c4485b962d8b6e9dd54fefb30	d7f6884ae0ae6e406ec3500fcde16e8f51642460	165415f700b0c77fa1f8db6198f48582639adf78
14	torvalds/linux	kernel/acct.c	2005-09-14 12:12:20 UTC	d7f6884ae0ae6e406ec3500fcde16e8f51642460	165415f700b0c77fa1f8db6198f48582639adf78	905ec87e93bc9e01b15c60035cd6a50c636cbaef
15	torvalds/linux	kernel/acct.c	2005-09-14 12:19:08 UTC	165415f700b0c77fa1f8db6198f48582639adf78	905ec87e93bc9e01b15c60035cd6a50c636cbaef	dbaa9a9d2b37d838125fb7f2b9fdc5dc5fa4eaa9
16	torvalds/linux	kernel/acct.c	2005-09-14 12:57:30 UTC	905ec87e93bc9e01b15c60035cd6a50c636cbaef	dbaa9a9d2b37d838125fb7f2b9fdc5dc5fa4eaa9	4e0c1159d83a658d1ffba5bc3442f4ec4cadb436
17	torvalds/linux	kernel/acct.c	2005-09-25 03:14:45 UTC	dbaa9a9d2b37d838125fb7f2b9fdc5dc5fa4eaa9	4e0c1159d83a658d1ffba5bc3442f4ec4cadb436	4294621f41a85497019fae64341aa5351a1921b7